A BERT-BiLSTM-CRF Approach to News Named Entity Recognition for Chinese news

He Zhang 23020231154256[1]
Jialin Ye 23020231154249[1], **Ruilin Jiang 23020231154193**[2], **Wu Liu 23020231154153**[2]
[1]AI Class
[2]Information Class

Abstract

Named Entity Recognition (NER) is one of the hot research directions in natural language processing. Traditional methods cannot well extract the local features and contextual semantics of news text. Therefore, it is of great significance to study and implement the news text named entity recognition method in this paper. This paper uses BERT, BiLSTM and CRF, and designs and implements a named entity recognition system supporting Chinese news texts, which can recognize basic entities such as names of people, places and organizations in the news, and has certain new word adaptation ability. The method of recognising named entities in Chinese and English news is studied. In order to obtain better recognition results, a BERT pre-training model is used; To address the problem that RNN cannot handle long sequences and LSTM cannot consider contextual information, BiLSTM is introduced to enhance the feature extraction capability. Considering the relationship between adjacent labels, CRF is introduced to optimise the annotation ability of the model.

Introduction

In the realm of Named Entity Recognition (NER) for Chinese news, researchers have explored various methodologies to enhance the accuracy and efficiency of NER systems. This section provides an overview of the existing approaches, shedding light on the evolution of techniques that paved the way for our proposed BERT-BiLSTM-CRF model.

One noteworthy research avenue in Chinese NER is the utilization of pre-trained word embeddings. Early studies harnessed traditional word embeddings like Word2Vec and GloVe to capture semantic meanings of words. However, these approaches faced challenges in handling out-of-vocabulary words and capturing complex word relationships.

To address these limitations, researchers turned to neural network-based models. BiLSTM, a bidirectional variant of Long Short-Term Memory networks, became a popular choice due to its ability to capture contextual information from both left and right contexts. Studies like (Zhao et al. 2019) demonstrated the effectiveness of BiLSTM in capturing long-range dependencies, improving the accuracy of Chinese NER systems.

The advent of transformer-based models marked a significant milestone in NER research. BERT (Bidirectional Encoder Representations from Transformers), introduced by (Devlin et al. 2019), captured intricate word relationships and contextual information effectively. Researchers extended BERT's capabilities by integrating it with sequence labeling models like CRF (Conditional Random Fields) . Huang and Xu showcased the synergy between BERT and CRF, resulting in improved NER performance, particularly for Chinese text(Zhang and Yang 2018).

Our work builds upon these foundations by proposing a BERT-BiLSTM-CRF approach for Chinese NER in news articles. We leverage the pre-trained BERT model to extract contextualized word embeddings, enhancing the model's understanding of complex language structures. BiLSTM layers capture intricate dependencies, and the CRF layer ensures coherent labeling sequences.

Related Work

News named entity identification refers to labelling certain text entities with specific meanings in news texts. Academically, named entity recognition can generally classify specific entities into three main categories and seven subcategories.

The research history of named entity recognition is broadly characterised by early rule and dictionary methods, traditional machine learning methods, deep learning methods, and the current trend of fusing deep learning with multiple models.

Early rule-based dictionary methods. In the early stages, researchers commonly employed rule-based and dictionary-based approaches for named entity recognition. This approach relied on manually crafted rule templates and primarily utilized pattern matching for identification. Rule templates could be devised based on criteria such as core words, keywords, punctuation, and then matched and recognized with specific text.When employing rule-based and dictionary-based methods, it is necessary to assign specific weights to the rules. While this method achieves high precision for specific text, its practicality is limited. It heavily depends on experts manually creating entity recognition libraries. Different languages possess distinct grammatical structures and usage rules, making this method challenging to achieve universality and compatibility. Error rates are also

relatively high, and the cost of usage is significant, requiring the development of corresponding recognition libraries for various language texts.

Methods based on statistical machine learning. With the continuous advancement of technology, methods based on statistical machine learning have emerged. Statistical methods can address the shortcomings of labor-intensive dictionary-based approaches . The introduction of this approach represents progress in named entity recognition methods, including Hidden Markov Models (Eddy 1996), Maximum Entropy Models (Kapur 1989), Support Vector Machines, Conditional Random Fields, and others. Named entity recognition in the field of machine learning can be regarded as a sequence labeling problem. By training models on large datasets, corresponding weights are obtained to assist researchers in entity recognition, thereby overcoming the labor-intensive nature of template-based methods.

Methods based on deep learning. Since Hinton first proposed deep learning models , numerous researchers have achieved remarkable results using this approach. Currently, the application of deep learning has led to a more profound development in news named entity recognition.One of the most representative approaches is the use of neural networks. This method exhibits excellent adaptability. Typically, researchers employ recurrent neural networks and convolutional neural networks to tackle named entity recognition tasks. In deep learning, news named entity recognition no longer solely relies on corpora and rule templates manually created by relevant scholars. Instead, text is tokenized and transformed into word vectors, significantly enhancing the universality and practicality of algorithms and models, optimizing performance, and reducing resource expenditure and waste.Researchers use convolutional neural networks to address named entity recognition problems and then combine conditional random fields with convolutional neural networks to further enhance model performance and precision. However, convolutional neural networks are feedforward neural networks, and they can be somewhat limited when dealing with tasks that involve close front-to-back relationships like named entity recognition. Therefore, an algorithm that can retain historical information is needed, and recurrent neural networks can be employed to utilize contextual relationships between texts for entity recognition, significantly improving accuracy. Nevertheless, recurrent neural networks also have their limitations. Lample et al. applied Long Short-Term Memory (LSTM) networks to named entity recognition tasks, overcoming the drawbacks of recurrent neural networks .

Hybrid methods. Most contemporary news named entity recognition work doesn't rely on a single method but instead combines multiple approaches to seek the best recognition performance. On top of this, various named entity recognition pre-trained language models, such as Google's BERT model, have emerged like mushrooms after the rain. These models have provided a more convenient and accurate solution for news named entity recognition. Currently, models that combine deep learning and statistical learning methods are the mainstream solutions for this work, such as the integration of pre-trained language models like BERT, Bidirectional Long Short-Term Memory (BiLSTM), and conditional random fields, among others. This article adopts such an approach.

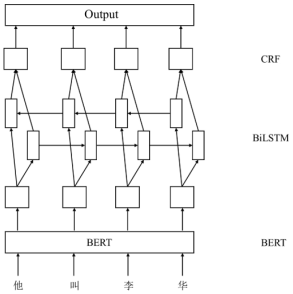

Figure 1: structure of BERT-BiLSTM-CRF model

Proposed Solution

Named Entity Recognition Models for Chinese News

The main focus of this section is to research and implement a named entity recognition model for Chinese news texts. The model consists of three main layers: BERT layer, BiLSTM layer, and CRF layer. The overall structure of the model in this section is illustrated in Figure 1.

The model operates by preprocessing the original Chinese news text to obtain an input sequence. This processed sequence is then fed into the BERT layer, which outputs vectors containing semantic information of the Chinese news text. The resulting vector is then passed to the BiLSTM layer, which captures contextual semantics from both the forward and backward directions, producing new outputs. The output from the BiLSTM layer is then input into the CRF layer, which considers the relationships between adjacent labels. After processing, the model obtains the final label results.

News Named Entity Recognition BERT Layer

In English news named entity recognition models, the main role of the BERT layer is to obtain vectors that contain the semantic information of the text.

Therefore, Google has researched and proposed a new model called BERT, which is a pre-trained language model for natural language processing. BERT is a bi-directional language model based on Transformers and can encode characters. Its central idea is to pre-train on a large amount of unlabeled data using an unsupervised approach, resulting in a highly adaptable language model. The BERT model can handle more than 11 different NLP tasks and has achieved good results in sentiment analysis, question answering, text prediction, named entity recognition, and other tasks.

The input part of BERT can be understood as a stack of multiple layers of Transformer encoders. It consists of the following three layers:

1. Token Embeddings: This layer extracts word roots from the pairs of news text sentences, tokenizes them, and marks the start of sentence pairs with $\langle CLS \rangle$. It also adds

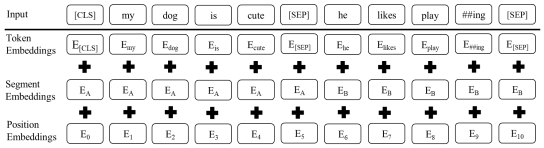

Figure 2: Input of the BERT Model

$\langle SEP \rangle$ between the two sentences to differentiate and annotate them, capturing the semantic information of the news text. This vector is automatically generated during model training to mark the beginning and end of the news text.

2. Segment Embeddings: This layer differentiates between the preceding and following sentences in the pairs of news text sentences to avoid confusion.

3. Position Embeddings: This layer represents the positions of the words in the news text sentences and can be annotated with different numbers.

The results of these three layers are combined to form the input of the BERT model, as shown in Figure 2.

Chinese News Named Entity Recognition BiLSTM Layer

In Chinese news named entity recognition models, the BiLSTM layer is responsible for extracting semantic information from Chinese news text in both left-to-right and right-to-left directions. The outputs of the forward and backward networks are combined to obtain a sequence with bidirectional semantics. After normalization, the label score matrix is obtained as the output of the BiLSTM layer.

In Chinese news named entity recognition tasks, the impact of different language environments on entity recognition is significant. Traditional RNNs have several drawbacks, such as gradient explosion when dealing with longer news character sequences and the inability to extract contextual semantics from both directions. Therefore, this paper adopts BiLSTM as the neural network model for this layer.

In Chinese news named entity recognition neural network models, the inputs of this layer are generally represented in vector form. Word vectors have relatively low dimensions and are dense. After the data enters the model, it undergoes various parameter transformations and is ultimately represented as a vector that captures the language environment and semantics of the current sentence. In order to achieve better and more accurate language prediction, it is generally desired for the positions of the vectors to be closer.

After the original news input sequence undergoes encoding transformation by the BERT layer, it produces vectors that contain contextual semantics. These vectors are then used as inputs to the BiLSTM layer. Compared to a single-layer LSTM model, BiLSTM can extract features from both directions of the Chinese news input sequence. Specifically, assuming the input is x and the output is y, the first forward LSTM layer extracts forward information from the sequence produced by the BERT layer, representing it as a vector. The second layer processes the text from the opposite direction and represents it as a reverse vector. Finally, the two vectors are merged to generate a total vector, which serves as the output y of this layer. The output is then processed through normalization to obtain the predicted label scores for the current news characters.

Working Principle of LSTM Networks In Chinese news named entity recognition models, the workflow of LSTM can be roughly divided into the following steps:

In the LSTM (Long Short-Term Memory) model, the forget gate is initially used to decide whether to discard information from the previous state when entering a new state. This decision, based on the previous state's output and current input, involves selectively forgetting (0) or retaining (1) information through calculated transformations, allowing the network to focus on relevant context.

Next, the input gate determines which information should be stored in the current state. It employs sigmoid transformation on the previous state's output and the current input, creating a reserved state vector using the tanh function. This vector is temporarily stored in the current state.

The state replacement step involves two multiplications and one addition. The first multiplication occurs between the previous state and the forget gate, while the second multiplication involves the reserved vector and the transformed retained input. Combining the results yields a new candidate state.

In the final output step, the candidate state undergoes multiplication with the output gate (after applying the sigmoid function). This multiplication determines the portion of the state to be outputted, completing one iteration of the LSTM process. This iterative process forms the overall structure of the LSTM model.

Structure of Bidirectional Long Short-Term Memory (BiLSTM) The LSTM model is effective in extracting feature information from long sequential inputs, but its limitation lies in its ability to work in only one direction. In different contexts, semantics can vary significantly, such as cases involving derogatory or complimentary language. This requires considering the problem from the opposite direction as well. Therefore, BiLSTM composed of two LSTM layers is used to address this issue.

A bidirectional LSTM is built upon the basic LSTM structure and consists of two LSTMs with opposite directions. One LSTM extracts features from the forward sequence, while the other LSTM extracts features from the backward sequence. A single LSTM can only capture information in one direction, but when two layers are used together, bidirectional information can be obtained, allowing the model to have a more comprehensive understanding of the contextual semantics.

Operation of BiLSTM In the Chinese news named entity recognition model, for an input sequence, it is fed into two separate memory networks in opposite directions. One network processes the sequence in the forward direction, while the other network processes it in the backward direction. After obtaining the output vectors separately, the two vectors are concatenated to produce the final output. At each time step, the bidirectional LSTM generates both forward and backward outputs. These outputs are then combined us-

ing concatenation or weighted averaging to form the final result. Over multiple time steps, a multidimensional result is obtained. This completes one iteration of the workflow, allowing the model to capture language context information from both directions, thus considering a more comprehensive and realistic context.

Chinese News Named Entity Recognition CRF Layer

The input of the CRF layer is the label score matrix obtained from the output of the BiLSTM layer. After passing through the BiLSTM layer, an output sequence is obtained that contains the semantic information of the Chinese news. However, this semantic information is partial, as it only considers the mutual relationships between individual characters and does not capture the associations between Chinese news entity labels. For example, in the BIO tagging scheme, the label following B-LOC should be I-LOC. But if only BiLSTM is used, it is possible to have B-LOC followed immediately by B-ORG, which is incorrect. Therefore, this paper introduces Conditional Random Fields (CRF) to address and optimize this problem.

CRF is an undirected graphical model that provides a good approach for predicting label sequences in natural language processing tasks. The key to CRF is conditional probability, which allows obtaining the desired label prediction information given known data.

Bidirectional Long Short-Term Memory networks can capture contextual semantic information of news text from both the forward and backward directions, but they cannot capture the relationships between labels. CRF, on the other hand, is essentially a statistical method and is currently the most commonly used approach in news text named entity recognition. CRF has many advantages. Compared to BiLSTM, CRF can fully consider contextual information and incorporate the connections between adjacent news entity labels, leading to more accurate predictions. CRF can add constraints to the predicted label results, such as requiring I or O to follow B, or specifying that named entities should start with B. These constraints can improve the accuracy of news named entity recognition.

Unlike the previous layer, CRF pays more attention to the overall linguistic context of the news text. It is a global model that trains based on the relationships between adjacent labels, giving it an advantage over BiLSTM in this aspect. In the training process, this paper successfully obtained the weight parameters of the corresponding entity labels for each character vector. The CRF layer abstracts the sequence into paths and calculates the scores of possible paths starting from the beginning. When transitioning from one label to another, the transition score between them is also considered. After obtaining the scores, normalization is applied using a function to represent the label probabilities of news named entities. The label with the highest probability is selected for output, resulting in the final result.

CRF is a probabilistic graphical model, as shown in Figure 3. It is used for modeling and processing sequence labeling and structured prediction problems. Each point in the model graph represents a labeling sequence and an observation sequence, which can be seen as random variables. The entire model can be considered as a whole composed of many points, where the probability of each point depends on the conditional probability with the previous and next nodes. When training the Chinese news named entity recognition model, CRF calculates the probabilities of all possible label sequences and selects the one with the highest probability as the output.

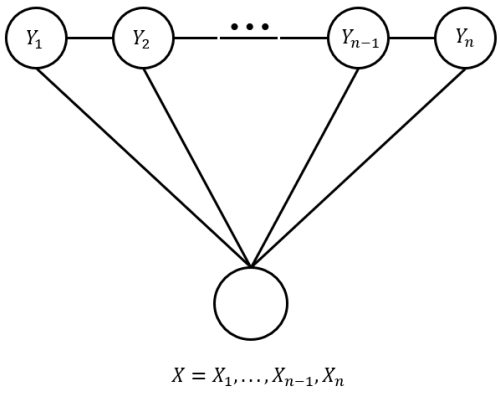

Figure 3: Illustation of CFR

Experiments

Processing News Text Dataset

In named entity recognition (NER), the initial step involves processing the raw corpus, where the training and recognition dataset includes labels and segmented entities. Various labeling schemes, such as BIO and BIOES, help identify the start and end of named entities, with labels comprising a special letter indicating word position and a category label denoting entity type.

In BIO, ”B” signifies the start, ”I” the middle, and ”O” indicates no entity association. In BMES, ”B” is the start, ”M” denotes middle characters, ”E” marks the end, and ”S” signifies single-character entities.

Common NER labels include person, location, and organization names. Chinese news introduces specific labels like person name (PER/NAME), location (LOC), organization (ORG), nationality (CONT), race (RACE), education (EDU), and title (TITLE).

Chinese NER faces challenges with the BIO format, unable to clearly distinguish middle and end parts of entities. The BMES format is preferred, providing hierarchical information. Adapting People’s Daily data from BIO to BMES requires manual code conversion, where ”B” and ”O” remain unchanged, ”I” depends on the next label, changing to ”E” or ”M” accordingly. Looping through the entire file achieves the desired labeling, completing data cleaning.

In addition, to enhance the system’s recognition capability, additional data containing other entities was added to the original People’s Daily dataset with specific labels. The added data was also annotated using the BMES method.

Training Parameters	Values
train_batch_size	16
train_epochs	3
dropout	0.3
lr	3e-5

Table 1: BERT-BiLSTM-CRF Model Parameter

Finally, after completing the data cleaning process, the dataset is divided into training, validation, and testing sets. Taking the Chinese dataset as an example, the number of entities in the training set is around 20,000, and the validation and testing sets each contain around 2,000 entities, ensuring an appropriate ratio among the three sets.

Evaluation metrics

Generally speaking, a well-performed named entity recognition system should meet the following criteria: correct entity boundaries, accurate position information and word segmentation, and correct entity types. To meet the first two criteria, it is necessary to correctly identify the starting and ending positions of entities. For example, the entities "Beijing" and "John" must be separated. If they are recognized as "Beijing John," it indicates a word segmentation error. Entity type is a direct criterion for measuring the success of named entity recognition, referring to the identified entity categories.

Named entity recognition is a classification task. In order to quantitatively analyze and study the results of named entity recognition, scholars have proposed the following three metrics for evaluation: Precision (P), Recall (R), and F1-Score.

Experimental Study

In terms of named entity recognition in Chinese news, the experimental parameter settings are shown in Table 1.

Among them, train_batch_size refers to the number of data samples inputted to the model in one batch. Train_epochs denotes the total number of times the model completes training on the training set. Dropout is used to prevent overfitting of the model. Lr represents the learning rate, which determines the step size at each iteration during the training process. In this experiment, the BERT-BiLSTM-CRF model was trained on a Chinese news dataset. First, the data was split and processed to transform it into the format required by BERT, which extracts semantic information from the news text. Then, the BiLSTM model was used to capture contextual semantics from both directions. The CRF layer was employed to consider the relationships between adjacent labels. After processing, the final predicted labels were outputted, and the experimental process was recorded and visualized.

Below is the training loss plot of this model on the Chinese dataset, as shown in Figure 4. By analyzing the loss plot, it can be observed that the model eventually converges.

After achieving satisfactory training results, the model was further evaluated on the test dataset, and the results are shown in Table 2.

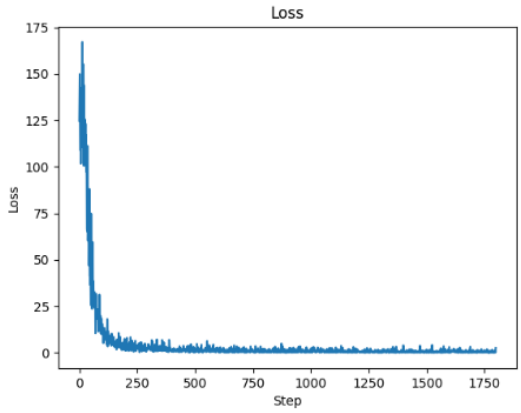

Figure 4: the loss plot of the BERT-BiLSTM-CRF model

Model Name	Precision	Recall	F1-Score
BERT-BiLSTM-CRF	93%	88%	91%

Table 2: BERT-BiLSTM-CRF Model Parameter

Conclusion

This study primarily focuses on the named entity recognition (NER) of Chinese news texts. A news NER system was designed and implemented, and training was conducted on a Chinese news dataset, achieving satisfactory results. The following is a summary of the work conducted in this paper:

1. Investigated the NER methods for Chinese news texts based on BERT, BiLSTM, and CRF.
2. Designed and implemented a news NER system that supports Chinese named entity recognition. The system automatically selects the appropriate model for recognition based on the input text language and has the capability to adapt to new words.
3. Conducted experimental verification on Chinese news datasets. The evaluation metrics used were precision, recall, and F1 score. In terms of Chinese news NER, the model achieved an F1 score of over 90

Overall, this study contributes to the development of NER methods for Chinese news texts and presents a practical news NER system with promising performance.

References

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Eddy, S. R. 1996. Hidden Markov models. *Current Opinion in Structural Biology*, 6(3): 361–365.

Kapur, J. N. 1989. *Maximum-entropy models in science and engineering*. John Wiley & Sons.

Zhang, Y.; and Yang, J. 2018. Chinese NER Using Lattice LSTM. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1554–1564. Melbourne, Australia: Association for Computational Linguistics.

Zhao, Z.; Chen, Z.; Liu, J.; Huang, Y.; Gao, X.; Di, F.; Li, L.; and Ji, X. 2019. Chinese named entity recognition in power domain based on Bi-LSTM-CRF. In *Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition*, 176–180.